Does the Order of Item Difficulty of the Addenbrooke's Cognitive Examination Add Anything to Subdomain Scores in the Clinical Assessment of Dementia

EXTRA Dementia and Geriatric Cognitive Disorders

Dement Geriatr Cogn Disord Extra 2015;5:155–169
DOI: 10.1159/000375364
Published online: April 15, 2015

1664–5464/15/0051–0155$39.50/0
www.karger.com/dee

Karger Open access

Original Research Article

Does the Order of Item Difficulty of the Addenbrooke's Cognitive Examination Add Anything to Subdomain Scores in the Clinical Assessment of Dementia?

Sarah McGrory[a, b] John M. Starr[a, c, d] Susan D. Shenkin[c, d]
Elizabeth J. Austin[b] John R. Hodges[e, f]

[a]Alzheimer Scotland Dementia Research Centre, [b]Department of Psychology, [c]Geriatric Medicine Unit, and [d]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK; [e]Neuroscience Research Australia, and [f]School of Medical Sciences, University of New South Wales, Sydney, N.S.W., Australia

Key Words

Mokken scaling · Hierarchical scales · Item response theory · Dementia · Addenbrooke's Cognitive Examination

Abstract

Background: The Addenbrooke's Cognitive Examination (ACE) is used to measure cognition across a range of domains in dementia. Identifying the order in which cognitive decline occurs across items, and whether this varies between dementia aetiologies could add more information to subdomain scores. ***Method:*** ACE-Revised data from 350 patients were split into three groups: Alzheimer's type (n = 131), predominantly frontal (n = 119) and other frontotemporal lobe degenerative disorders (n = 100). Results of factor analysis and Mokken scaling analysis were compared. ***Results:*** Principal component analysis revealed one factor for each group. Confirmatory factor analysis found that the one-factor model fit two samples poorly. Mokken analyses revealed different item ordering in terms of difficulty for each group. ***Conclusion:*** The different patterns for each diagnostic group could aid in the separation of these different types of dementia.

Introduction

Cognitive measures are commonly used to screen for dementia as well as assessing severity and monitoring disease progression. Often underlying these tests is the assumption that cognition deteriorates along a fixed course of decline on a single cognitive trait (i.e. total

Sarah McGrory
Alzheimer Scotland Dementia Research Centre
University of Edinburgh
7 George Square, Edinburgh EH8 9JZ (UK)
E-Mail s.mcgrory@sms.ed.ac.uk

test scores are considered meaningful in themselves) and that the impairment and severity can be measured when a patient is unable to respond correctly to certain cognitive challenges [1]. Looking at total and subdomain scores may lead to important information being neglected. For example, two different individuals achieving the same score on a cognitive measure may have reached this score by missing different combinations of items. Using the summed score as a measure of cognition fails to take into account the information embedded in the specific pattern of scores. Items may differ in several ways. Different items on a scale may be unequally related to the construct of cognitive impairment. Additionally, test items are likely to differ in terms of difficulty – how difficult an individual finds it to respond correctly to an item [2].

The Addenbrooke's Cognitive Examination (ACE) [3] was originally developed to provide a brief test that would be both sensitive to the initial symptoms of dementia and capable of discriminating different types of dementia including Alzheimer's disease (AD) and frontotemporal dementia (FTD) [3]. The ACE and the revised version (ACE-R) [4] encompass tests of attention/orientation, memory, language, visuospatial abilities and executive function. They also incorporate the Mini-Mental State Examination [5], so this score may also be produced. The ACE is relatively quick to administer (approximately 15 min) and has good sensitivity and specificity for identifying dementia [3]. While modifications to the ACE have been made to address the original scale's weaknesses, there have been no examinations of the item properties or 'hierarchical' structure of either the original ACE or its successor the ACE-R using item response theory (IRT) [6] methods. A hierarchical scale in the context of IRT implies that all scale items are ordered relative to each other in terms of their level of difficulty (i.e. the ease with which an item is responded to) and that all are ordered along the latent trait being measured. All references to hierarchical scales in this paper will refer to this IRT hierarchical ordering by difficulty.

Factor analysis can be used to investigate the relationship between ACE-R items and the total score. While this method offers some insight into the dimensionality of the ACE-R, IRT can provide further insight into the item properties and how they function in relation to the other items within the scale. This item level analysis can be applied to determine the items for a hierarchy of item difficulty.

The interpretation of the ACE-R and other cognitive measures would be greatly improved if the ordering of the difficulty of the cognitive tasks (items) was similar for patients at different stages of dementia. When the ordering of the items by mean scores is the same across different values of the latent construct, it can be said that items conform to a hierarchical scale with invariant item ordering (IIO). Fundamentally, IIO means that the items in a scale 'have the same order with respect to difficulty or attractiveness for all respondents' [7, p. 578]. IIO is a very important property as once this has been established, the item ordering within the scale in question will be the same for the population of interest, along with any subgroup of the population. IIO can facilitate diagnosing dementia [8]. For example, an IIO hierarchy detailing the expected trajectory of decline in AD may differ from an IIO hierarchy of decline in semantic dementia (SD). In this way, IIO hierarchies can be used to identify distinctive profiles of cognitive dysfunction which can serve as an adjunct to diagnosis. IIO can also facilitate the comparison of patients with respect to their degree of cognitive decline, for example a patient experiencing problems with one of the least difficult items in the hierarchy would be considered more severely impaired than a patient only experiencing a problem with one of the most difficult items in the hierarchy. IIO hierarchies can also be useful in the detection of unexpected score patterns [9] and in characterising differences among subgroups and different forms of dementia.

Mokken scaling analysis [9, 10], based on IRT principles, is commonly applied to determine whether hierarchical scales meeting IIO criteria exist within data. This method has been more

frequently applied to dichotomous items within scales. However, examining polytomous scales (i.e. scales with more than two response options, for example 'strongly disagree', 'disagree', 'agree', 'strongly agree' or an item with a score range of 0–3) for IIO has recently become possible [11, 12].

The aim of the present study is to determine whether the ACE-R has hierarchical properties with IIO and to compare these findings with factor analysis using structural equation modelling to determine whether a hierarchy of item difficulty can add to the information provided by the subdomain scores.

Methods

Participants

A sample of 350 patients was sourced from the specialist multidisciplinary tertiary referral centre FRONTIER (the Frontotemporal Dementia Research Group) at Neuroscience Research Australia (NeuRA), Sydney. Patients meeting current clinical diagnostic criteria for behavioural variant FTD (bv-FTD) [13], AD [14], logopenic progressive aphasia (LPA) [15], motor neurone disease (MND) [16], progressive non-fluent aphasia (PNFA) or SD [17] were recruited through FRONTIER. Diagnosis was established by consensus among neurologist, neuropsychologist and occupational therapist, based on extensive clinical assessments, cognitive assessment and evidence of atrophy on structural MRI brain scans. All patients provided informed consent for the study, and dual consent was obtained from the carer in some cases. Patients underwent clinical, neuropsychological, behavioural and imaging assessment between 2007 and 2011. Data from patients with complete itemised ACE-R data (n = 350) were included in the analysis.

The sample was very diagnostically heterogeneous, and in an attempt to limit the effects of this heterogeneity, the sample was divided into three groups: AD type: AD and LPA (n = 131); predominantly frontal dementia: bv-FTD and FTD-MND (n = 119); other frontotemporal lobe degenerative disorders, other frontotemporal lobe degenerative disorders temporal: SD and PNFA (n = 100).

Measures

The ACE-R [4] comprises 26 items, is scored out of 100 and includes items assessing 5 cognitive domains: attention/orientation (18 points), memory (26 points), fluency (14 points), language (26 points) and visuospatial (16 points). The total ACE-R score is created by the addition of all item scores across all domains.

The mean for each ACE-R item score was divided by the maximum number of points available for that item to equate scores for comparison (i.e. equal weighting of items even though items can contribute different weighted values to the summed total score), giving an item score with minimum 0 and maximum 1. For example, the mean score of 2.5 for 'memory retrograde' for the predominantly frontal group was divided by 4 (the maximum number of points available on this item) to give a new 'overall' mean score of 0.625. These equated mean item scores were used for the analyses.

Although the rescoring of 'naming (10 items)' potentially removes some important variation in response, this was minimised by collapsing the item responses at the bottom end of the range since the prevalence of responses in the lowest category is very low (n = 34, 9.7%).

Factor Analyses

To identify the underlying factor structure, an exploratory principal component analysis (PCA) was performed on the subdomain scores for each of the diagnostic groups using the

IBM SPSS, version 19. Inspection of scree plots and the Kaiser criterion of eigenvalues >1 were used to decide on the number of components to extract.

The final factor solution derived from the PCA was entered into AMOS and converted to a simple structure confirmatory factor analysis (CFA) model, in which one latent variable explained the covariance in the five subdomains. CFA was performed on the emergent factor structure to evaluate whether the PCA model fit the data well. The Comparative Fit Index (CFI) [18] and the root mean square error of approximation (RMSEA) [19] were used to estimate the model fit. The following rules of thumb with regard to model fit were used: CFI <0.90 indicates a poor fit, 0.90 < CFI < 0.95 indicates a reasonable model fit, and CFI >0.95 indicates a good model fit; RMSEA >0.10 indicates a poor fit, 0.5 < RMSEA < 0.10 indicates a reasonable model fit, and RMSEA <0.5 indicates a good fit [20]. All confirmatory analyses were conducted with AMOS 19.0 [21].

Mokken Scaling Analysis

To determine whether the ACE-R conforms to a hierarchical scale and, if so, how this hierarchy relates to the factor structure, Mokken scaling analysis was carried out. Data were analysed using the Mokken scaling analysis package in the public domain software R, which tests the assumptions of both Mokken models: the monotone homogeneity model and the double monotonicity model [22].

The 26 items of the ACE-R were analysed. Some of these items are a composite of several embedded questions such as 'orientation in time' on which a patient receives a score from 0 to 5 based on their ability to correctly identify the correct day, date, month, year and season. Mokken scaling permits polytomous data, but as only the total score from these items (i.e. a score out of 5 for 'orientation in time') was reported, the embedded items (e.g. 'what is the month?') could not be isolated for analysis. Instead, Mokken scaling was performed on the polytomous composite item score.

Mokken scaling is a non-parametric method for establishing whether hierarchical scales exist. Non-parametric methods are attractive for these data for several reasons. The ordinal measurement of respondents is guaranteed once the model applies to the data, and crucially non-parametric models are less restrictive requiring minimal assumptions. As the ACE-R data are not normally distributed, applying Mokken scaling techniques avoids forcing the data into a structure they do not have. A key advantage of non-parametric IRT methods over more commonly used parametric models such as those by Rasch [23] is that they have less strict assumptions regarding the non-linear behaviour of response probabilities in comparison to parametric IRT methods [9]. Within the Mokken scaling framework, no assumption is made about the shape of the relationship between the response to an item and the score on the latent trait – best described as the item response function (IRF) – aside from requiring IRFs to be monotonic and non-intersecting. Monotonicity in terms of the IRF implies that the higher the latent trait value, the higher the probability of a correct response to an item measuring the latent trait. In this way, the IRF is an increasing function of the latent trait [9]. Non-intersection or IIO is a property of IRFs, whereby the IRFs for the scores of items do not intersect. The strong assumptions of the more restrictive parametric methods increase the likelihood of potentially useful items being rejected due to the shape of their IRFs.

Mokken scaling procedures provide several parameters for determining whether the data conform to Mokken scales. While Mokken scaling is considered as a probabilistic reworking of the deterministic Guttman scaling [24], the strength of Mokken scales is determined based on the number of Guttman errors [25]. A Guttman error can be observed when the relative response to a pair of items is not in the expected direction [26]. The fewer the Guttman errors, the stronger the Mokken scale is considered [9]. Loevinger's *H*, a measure of the strength and quality of a Mokken scale, is used to indicate the extent of Guttman errors and as such is an

expression of the degree to which the items consistently appear in the same relative order and justifies their use in forming a unidimensional latent variable [9]. The overall H for the scale as a whole along with a coefficient (H_i) for each of the individual items within the scale is calculated. The item scalability coefficient (H_i) is a reflection of item discrimination with higher values reflecting a greater ability of the item to differentiate between different levels of latent trait. Items with low H_i values (e.g. <0.3) indicating poor discrimination are generally excluded to ensure that all scale items are at least moderately discriminatory [27]. Similarly, $H = 0.3$ is the minimum value for a Mokken scale with higher values reflecting greater strength of ordering and fewer violations [28]. Violations, or Guttman errors, are defined as any deviations of the data from the expected ordering. For example, if there is an item *(i)* with a lower mean score than another item *(j)*, indicating that item *i* is a more 'difficult' item, then any time item *i* has a higher score than item *j*, an error has occurred [29].

Monotonicity can be examined by calling the function *check.monotonicity* in R. This function calculates the number of scaling violations, where the predicted order of an item pair is reversed, and summarises these in the output for inspection.

Another of Mokken's assumptions, local stochastic independence (LSI) of items, implies that a subject's response to one item in the test is not affected by his or her response to any other item in the scale. LSI implies that all systematic variation in responses to the items is exclusively caused by the variation of respondents over θ [30]. By nature, items belonging to the same scale have to co-vary to some degree [31], but LSI implies that this item covariance is due to the latent trait they all measure [9]. Methods for estimating stochastic dependence in polytomous items within the Mokken scaling procedure are in development but are currently unavailable [32].

These diagnostics and parameters are used to establish whether the assumptions of the monotone homogeneity model hold and are used to assess the fit of the data to Mokken's first level of analysis, the scalability of the items. Scalability measures the extent to which respondents can be reliably ordered on the level of latent trait by means of their summed total score [33]. However, it is important to note that while H can inform on whether a set of items form a unidimensional and monotonic scale, it is not sufficient to determine whether the items form a hierarchical scale [34].

Only Mokken's second level of analysis, the assessment of IIO, can confirm whether a set of items forms a hierarchical scale [34]. For dichotomous items, IIO is established by examining non-intersection of IRFs. Non-intersection in the case of polytomous items is established where there is no intersection in each of the steps between response categories. The relationship between these responses and the score on the latent trait is symbolised using item step response functions (ISRFs). In the case of an item scored from 0 to 5 where there are four steps between the five possible responses, there are four ISRFs. Non-intersection is established where there is no intersection in each of the steps between these response categories. However, the non-intersection of ISRFs does not imply IIO [34, 35]. Most software cannot be used to assess IIO for polytomous items as only the ISRFs within each item are analysed and not the items themselves. This key element of Mokken scaling is only currently possible using the 'mokken' package of R software using the function *check.iio* [12]. This function uses a backward selection procedure that starts with the items with the highest number of IIO violations and iteratively removes items until there are no significant violations of IIO remaining. Using this method, a diagnostic Htrans or H^T is used to establish the strength of IIO, similar to the heuristics of H, with H^T values >0.3 indicating a scale with IIO [36]. Visual inspection of item pair plots was also used to assess IIO. Item rest-score regression plots were visually inspected to identify item overlap or 'outlying' items: items located far away from the cluster of the other scale items. These items can cause artificially exaggerated IIO and can result in the misleading appearance of IIO [8].

Table 1. Demographic and cognitive scores in dementia groups

	AD type	Predominantly frontal	Other frontotemporal lobe degenerative disorders
Male	131 (55%)	119 (77%)	100 (68%)
Age, years	66.5±8.3	63.7±8.9	65.8±8.9
Education, years	12.9±3.4	12.3±3.3	12.0±3.6
MMSE	22.0±5.7	23.7±5.8	21.3±5.8
ACE-R	64.2±19.0	70.3±18.9	54.9±17.3

Values are expressed as mean ± standard deviation or numer with percentage in parenthesis. MMSE = Mini-Mental State Examination.

Results

A total of 350 participants (232 male, 118 female) with a mean age of 65.38 (standard deviation = 8.5) years, diagnosed with dementia were included in the analysis (table 1). The sample was diagnostically heterogeneous: bv-FTD (n = 96), AD (n = 88), SD (n = 61), LPA (n = 43), PNFA (n = 39) and FTD-MND (n = 23).

Demographic and Cognitive Scores in Dementia Groups

Equated ACE-R item scores (table 2) were used to designate item difficulty in Mokken scaling.

PCA Analysis

Visual inspection of scree plots and Kaiser's criterion were used to extract factors with an eigenvalue >1. Both methods suggested a single-factor structure with the extraction of one component with an eigenvalue >1 for the AD type, predominantly frontal, and patients diagnosed with other frontotemporal lobe degenerative disorders, explaining 65, 68 and 61% of the variance in the groups, respectively. The correlations between the extracted component and the ACE-R subdomains are similar across the three diagnostic groups, as shown in table 3.

CFA Analysis

This one-factor model derived from PCA was converted to a CFA model. CFA was performed to evaluate whether the PCA model fit the data well. Whereas PCA examined all variance, the CFA model examined the shared variance. This model fit the data well in the predominantly frontal group ($\chi^2 = 6.754$, d.f. = 5; CFI = 0.994; RMSEA = 0.054) but less successfully in the AD type groups ($\chi^2 = 18.405$, d.f. = 5; CFI = 0.957; RMSEA = 0.143) and other frontotemporal lobe degenerative disorders group ($\chi^2 = 40.327$, d.f. = 5; CFI = 0.841; RMSEA = 0.269). Based on the CFI and RMSEA, the one-factor model fitted the AD type and other frontotemporal lobe degenerative disorders groups poorly.

Mokken Scaling Analysis

AD Type

Mokken's scalability coefficients were examined to assess the unidimensionality of the items. The H_i values of 6 items were below the recommended threshold level (0.3) for retaining items. These items: 'draw overlapping pentagons', 'draw a cube', 'count dot arrays', 'follow written instruction', 'three-item recall', and 'repetition-no ifs, ands or buts' were removed. These low values suggest that the items have weak discriminatory power. There were no

Table 2. ACE-R items grouped by domain, means and total scores

Item	Domain	Label	Mean			Max.
			AD type	predominantly frontal	other frontotemporal lobe degenerative disorders	
1	Attention	Orientation in time	3.5	3.8	4.1	5
2		Orientation in geography	3.9	4.1	3.6	5
3		Three-item registration	2.7	2.9	2.6	3
4		Serial sevens	3.9	4.0	3.9	5
5	Memory	Three-item recall	1.1	1.5	1.1	3
6		Name and address learning	4.6	5.8	4.6	7
7		Memory retrograde	2.0	2.5	1.2	4
25		Name and address recall	1.6	3.1	1.6	7
26		Recognition	3.5	3.7	3.3	5
8	Fluency	Verbal fluency-letters	3.4	2.5	2.0	7
9		Verbal fluency-animals	2.5	2.5	1.3	7
10	Language	Written instruction	0.8	0.9	0.8	1
11		Syntactical comprehension	2.3	2.5	2.1	3
12		Write a sentence	0.8	0.8	0.7	1
13		Repetition of single multi-syllabic words	1.4	1.6	1.1	2
14		Repetition-above, beyond and below	0.7	0.8	0.6	1
15		Repetition-no ifs, ands or buts	0.4	0.4	0.3	1
16		Naming (pencil and watch)	1.7	1.9	1.3	2
17		Naming (10 items)	6.2	6.8	2.5	9[1]
18		Semantic comprehension	3.1	2.8	1.8	4
19		Word reading	0.6	0.7	0.2	1
20	Visuospatial	Draw overlapping pentagons	0.7	0.8	0.9	1
21		Draw a cube	1.2	1.5	1.7	2
22		Draw a clock	3.4	3.9	3.5	5
23		Count dot arrays	3.4	3.4	3.6	4
24		Identify fragmented letters	3.7	3.9	3.7	4

Max. = Maximum score for each ACE-R item. All scores of 0 (n = 34, 9.7%) were recoded as 1 to provide a range of 0–9 as MSA is unable to analyse items with scores >9.

[1] Maximum score for naming (10 items) = 10.

Table 3. Correlations between ACE-R subdomains and the component extracted from PCA

	Component 1		
	AD type	predominantly frontal dementia	other frontotemporal lobe degenerative disorders
Attention	0.855	0.865	0.827
Memory	0.857	0.861	0.815
Fluency	0.790	0.775	0.775
Language	0.838	0.842	0.842
Visuospatial	0.666	0.781	0.629

Subdomain value derived from addition of mean item scores within each domain.

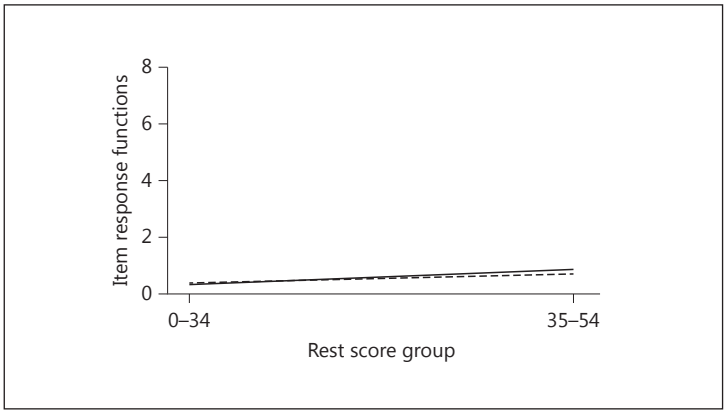

Fig. 1. Item pair plot demonstrating the intersection between 'repetition above, beyond and below' (solid line) and 'reading' (dashed line).

violations of monotonicity. Therefore, the remaining 20 items were deemed sufficiently homogenous to be unidimensional on the basis of the item scalability coefficients and *H* of 0.45.

Assessment of IIO resulted in 32 violations, 16 of which were significant. Starting with the item with the greatest violation, items were removed iteratively until no further violations remained. This process prompted the removal of a further 6 items ('identify fragmented letters', 'verbal fluency-animals', 'name and address learning', 'semantic comprehension', 'verbal fluency-letters', 'repetition of single multi-syllabic words'). The removal of these items resulted in 14 out of the original 26 items being retained in a moderately strong hierarchical Mokken scale [H = 0.44, standard error (SE) = 0.04] with IIO (H^T = 0.69). Inspection of item pair plots resulted in the exclusion of a further 3 items; 'repetition-above, beyond and below' and 'reading' were shown to intersect (fig. 1) and 'naming (10 items)' was identified as being located at some distance from the other items (fig. 2) which could be driving the high H^T value. The removal of these additional items left 11 items conforming to a moderate Mokken scale (H = 0.43, SE = 0.04) and lowered the strength of IIO (H^T = 0.52).

Discriminatory values of these items are presented in table 4 in the order of decreasing item scalability coefficients. SEs of scalability coefficients are also provided.

Table 5 presents the IIO items ordered according to their difficulty level (by mean score). These items have the same difficulty ordering irrespective of the value of the respondent's cognitive ability.

Predominantly Frontal Dementia

Four items were removed due to low H_i values; 'draw a cube', 'repetition-no ifs, ands or buts', 'repetition of single multi-syllabic words' and word reading'. There were no violations of monotonicity. The remaining 22 items were sufficiently homogenous to be considered unidimensional (H = 0.52).

There were 42 violations of IIO, 32 of which were significant. This process resulted in the removal of 6 items ('identify fragmented letters', 'three-item registration', 'syntactical comprehension', 'verbal fluency-animals', 'verbal fluency-letters', 'name and address recall'). Following the removal of these items, 16 items were retained in a strong Mokken scale (H = 0.52, SE = 0.05) with IIO (H^T = 0.82).

Some intersection was observed from visual inspection of the item pair plots. This warranted the further exclusion of 4 items: 'write a sentence', 'draw intersecting pentagons', 'repetition-above beyond and below' and 'follow written instruction-close eyes' (fig. 3). This left 12 items which conformed to a strong Mokken scale (H = 0.54, SE = 0.05) with a lowered

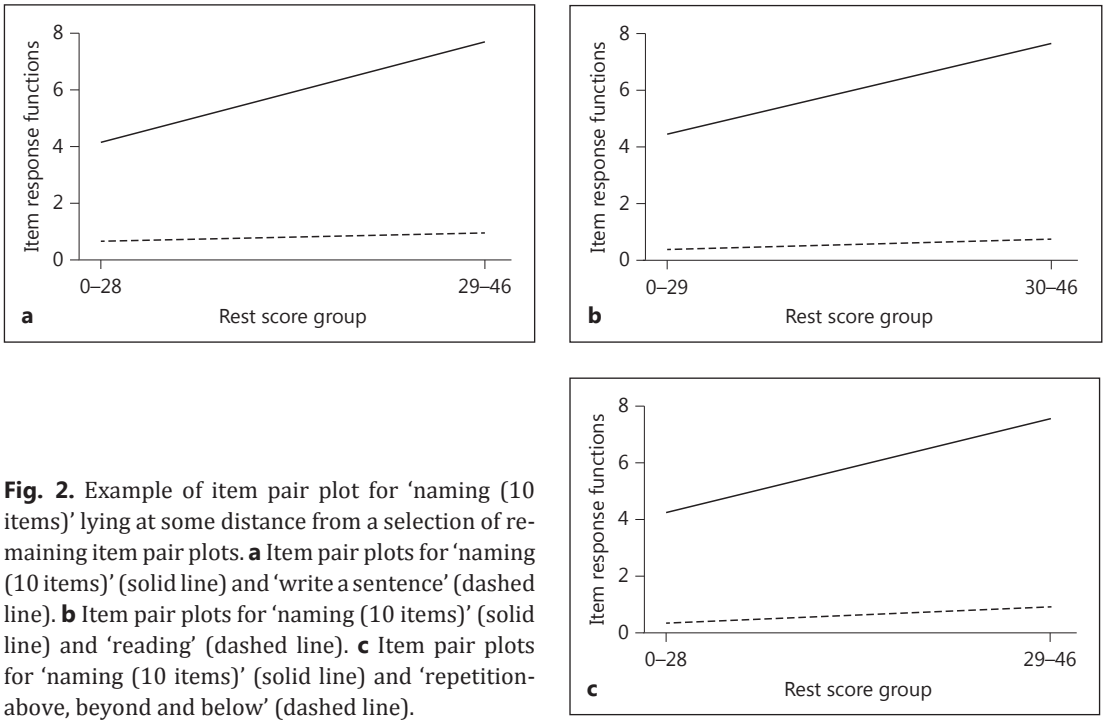

Fig. 2. Example of item pair plot for 'naming (10 items)' lying at some distance from a selection of remaining item pair plots. **a** Item pair plots for 'naming (10 items)' (solid line) and 'write a sentence' (dashed line). **b** Item pair plots for 'naming (10 items)' (solid line) and 'reading' (dashed line). **c** Item pair plots for 'naming (10 items)' (solid line) and 'repetition-above, beyond and below' (dashed line).

Table 4. Items listed in order of decreasing discrimination for each of the three groups

AD type			Predominantly frontal dementia			Other frontotemporal lobe degenerative disorders		
item	H_i	SE	item	H_i	SE	item	H_i	SE
Name and address recall	0.51	0.05	Draw a clock	0.60	0.05	Name and address recall	0.55	0.05
Naming (pencil and watch)	0.49	0.06	Name and address learning	0.58	0.05	Orientation in time	0.54	0.05
Orientation in geography	0.47	0.05	Recognition	0.58	0.05	Reading	0.53	0.09
Memory retrograde	0.46	0.05	Naming (pencil and watch)	0.57	0.07	Fluency-animal	0.52	0.05
Write a sentence	0.46	0.07	Orientation in geography	0.57	0.06	Recognition	0.51	0.05
Serial sevens	0.44	0.05	Serial sevens	0.56	0.06	Name and address learning	0.48	0.05
Three-item registration	0.43	0.08	Semantic comprehension	0.55	0.05	Three-item registration	0.45	0.08
Draw a clock	0.42	0.05	Orientation in time	0.51	0.06	Orientation in geography	0.42	0.06
Recognition	0.41	0.05	Naming (10 items)	0.50	0.06	Draw a clock	0.41	0.05
Syntactical comprehension	0.37	0.06	Memory retrograde	0.49	0.06	Write a sentence	0.40	0.08
Orientation in time	0.35	0.06	Three-item recall	0.48	0.06	Memory retrograde	0.40	0.07
			Count array of dots	0.38	0.07	Serial sevens	0.39	0.07
						Three-item recall	0.37	0.07
						Fluency-letter	0.34	0.06

strength of IIO (H^T = 0.72; see table 4 for item ordering by discrimination and table 5 for item ordering by difficulty).

Other Frontotemporal Lobe Degenerative Disorders

Ten items were removed due to low H_i values ('follow written instruction', 'repetition of single multi-syllabic words', 'repetition-above, beyond and below', 'repetition-no ifs, ands or buts', 'naming (10 items)', 'draw overlapping pentagons', 'draw a cube', 'count dot arrays', 'syntactical comprehension', 'semantic comprehension'). Again, there were no violations of monotonicity. The remaining 16 items were sufficiently homogenous to be considered unidimensional (H = 0.44).

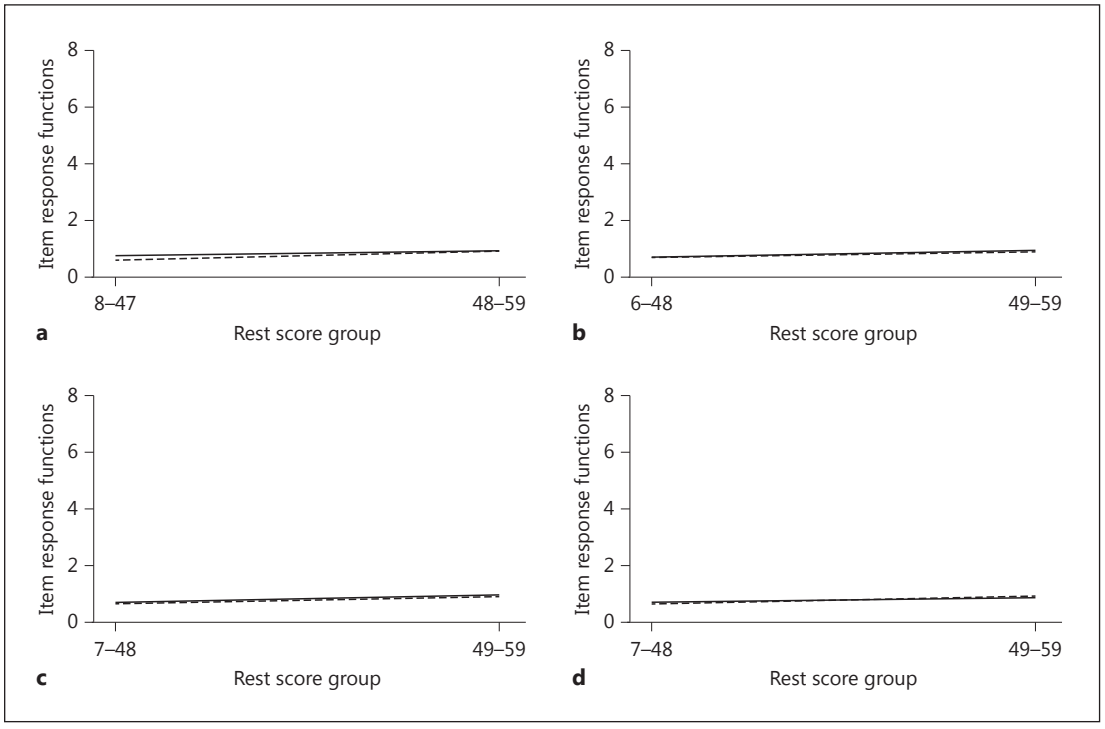

Fig. 3. Example of intersecting items from predominantly frontal dementia analysis. **a** Item pair plots for 'follow written instruction-close eyes' (solid line) and 'write a sentence' (dashed line). **b** Item pair plots for 'draw intersecting pentagons' (solid line) and 'repetition above, beyond and below' (dashed line). **c** Item pair plots for 'draw intersecting pentagons' (solid line) and 'write a sentence' (dashed line). **d** Item pair plots for 'repetition-above, beyond and below' (solid line) and 'write a sentence' (dashed line).

Table 5. IIO hierarchies with items ordered from most to least difficult in each diagnostic group

AD type		Predominantly frontal dementia		Other frontotemporal lobe degenerative disorders	
domain: item	mean	domain: item	mean	domain: item	mean
Memory: Name and address recall	0.22	Memory: Three-item recall	0.50	Language: Word reading	0.17
Memory: Retrograde	0.50	Memory: Retrograde	0.61	Fluency: Animal	0.19
Visuospatial: Draw a clock	0.68	Language: Semantic comprehension	0.70	Memory: Name and address recall	0.23
Memory: Recognition	0.69	Memory: Recognition	0.74	Fluency: Letter	0.29
Attention: Orientation in time	0.70	Language: Naming (10 items)	0.75	Memory: Retrograde	0.31
Attention: Serial sevens	0.77	Attention: Orientation in time	0.77	Memory: Three-item recall	0.36
Language: Syntactical comprehension	0.78	Visuospatial: Draw a clock	0.79	Memory: Recognition	0.65
Attention: Orientation in geography	0.79	Attention: Serial sevens	0.81	Memory: Name and address learning	0.66
Language: Write a sentence	0.84	Attention: Orientation in geography	0.82	Language: Write a sentence	0.69
Language: Naming (pencil and watch)	0.87	Memory: Name and address learning	0.83	Visuospatial: Draw a clock	0.71
Attention: Three-item registration	0.91	Visuospatial: Count dot arrays	0.85	Attention: Orientation in geography	0.71
		Language: Naming (pencil and watch)	0.92	Attention: Serial sevens	0.78
				Attention: Orientation in time	0.82
				Attention: Three-item registration	0.88
$H^T = 0.52$		$H^T = 0.72$		$H^T = 0.65$	

Mean = Mean ACE-R item score reflecting item difficulty; lower scores indicate greater item difficulty.

There were only two violations of IIO. This resulted in the exclusion of a further 2 items ['identify fragmented letters', 'naming (pencil and watch)']. Following the removal of these items, 14 items were retained in a moderate Mokken scale (H = 0.45, SE = 0.04) with IIO (H^T = 0.65). No further items were excluded following inspection of item plots.

Items of this IIO subset are presented in the order of discrimination (table 4) and difficulty (table 5).

Discussion

This study aimed to determine if hierarchies of ACE-R items meeting IIO criteria were present in three different samples consisting of different dementia diagnoses and to establish whether these hierarchies add anything to the subdomain scores and factor structure revealed by PCA. Mokken scaling analyses of the full 26 items of the scale for each of the three samples resulted in 11 items being retained in an IIO hierarchy in the AD type sample, 12 items in the predominantly frontal dementia sample and 14 items in the other frontotemporal lobe degenerative disorders sample.

The results of PCA did not indicate a difference between groups, with all groups being dominated by a large single component with similar item loadings. However, CFA analyses indicate that the structure of the one-factor model did not fit two of the three groups: AD type and other frontotemporal lobe degenerative disorders. Combining results of exploratory and confirmatory factor analyses and Mokken scaling analyses suggests that the factor structure of the ACE-R domains and the item ordering by difficulty within domains differ between diagnostic groups. These domain and item level differences could be applied to further differentiate different types of dementia by their associated ACE-R performance profiles.

In comparison with factor analysis, Mokken scaling has considerable theoretical and practical advantages [28]. Whereas factor analysis identifies groups of highly correlating items, Mokken scaling can illustrate the systematic order relationship between the items in a scale which improves construct validity [28]. Additionally, factor loadings disregard how item performance may differ across levels of the latent trait [37]. These advantages of Mokken scaling offer meaningful clinical implications. Establishing a formal hierarchy of item difficulty within a scale adds to the possibilities of interpretation and application. Hierarchical scales are appealing for their ease of use and scoring [38]. Responses to individual items, not just total scores, can provide an insight into a patient's level of ability based on the item's degree of difficulty [39]. For example, across the three diagnostic groups, a patient responding correctly to the 'memory retrograde' item is unlikely to have difficulty with any of the less difficult items. This insight enables quicker estimations of a patient's cognitive functioning and can facilitate adaptive testing, whereby only a selection of items, either from the more difficult or the less difficult range of the scale depending on the ability of the specific patient, is required for testing [40]. Tailoring tests to specific levels of ability can reduce testing time and stress and burden on patients.

Mokken scaling of dementia screening instruments can be used to assess whether the cognitive abilities are lost – or retained – hierarchically. Establishing whether these hierarchies differ across diagnostic groups can be useful in differential diagnosis. Although the comparison between the hierarchies is hampered by the lack of common items between the hierarchies, there are several notable differences in the ordering of item difficulty among the common items between the groups. Looking at the ordering of these items can provide some insight into the order of progressive decline in each group.

Practically, if we examine the item ordering of the following items: 'memory retrograde', 'orientation in time', 'orientation in geography', 'draw a clock' and 'write a sentence', some

different patterns emerge across the three groups: (a) if 'write a sentence' is less difficult than 'draw a clock', the patterns here suggest a diagnosis of AD or LPA is more likely than a diagnosis of SD or PNFA, (b) where 'write a sentence' is less difficult than 'orientation in time', the ordering here suggests AD type rather than the SD and PNFA group, (c) if 'memory recognition' is less difficult than 'draw a clock', these results indicate that a diagnosis of AD or LPA is most likely, and (d) comparing 'orientation in time' with 'orientation in geography', if the score for 'orientation in geography' is lower than that for 'orientation in time', the most likely diagnosis from those considered here is one of the other frontotemporal lobe degenerative disorders, either SD or PNFA. Since both 'orientation in time' and 'orientation in geography' scores contribute to the attention and orientation subscale of the ACE-R, conventional comparisons between subscales are insensitive to this difference between diagnostic groups.

Mokken analysis demonstrates that the IIO hierarchies present a more mixed profile than what can be observed from mean subdomain scores with some items within the domains being more difficult than others; for example, in the predominantly frontal group, 'name and address learning' is less difficult than the other memory items. There is a wide spread of the language items in terms of difficulty with language items found among the most and least difficult items. These differences are not captured using domain subscores.

Some limitations of this study are important to consider when interpreting the results. Fundamentally, the sample size of each of the groups analysed here is relatively small, particularly the other frontotemporal lobe degenerative disorders group (n = 100). This sample is very small for all analyses in this study. It would have been interesting to compare results from item level CFA and item level Mokken scaling analysis; however, the limited numbers available for analysis restricted the level of CFA permitted. This would be an interesting inclusion in a future study. The decision to include this small sample was made as the diagnoses within this group – SD and PNFA – are relatively uncommon, and accordingly large numbers of data are difficult to obtain. Therefore, while data from this group were included, results obtained must be interpreted with caution. Results from this sample and all analyses here require replication in a larger sample.

There has been little research providing minimum sample size requirements for Mokken scaling until recently. A simulation study investigating adequate sample sizes for Mokken scaling determined that the strength of item scalability coefficients, H_i, is inversely proportional to sample size, which serves as a good indicator of adequate sample size [32]. In this study's analysis of the smallest sample – other frontotemporal lobe degenerative disorders – 10 items were excluded due to low H_i values. It is very likely that this is a consequence of the small number of participants in this sample. A larger sample size may have resulted in the inclusion of a greater number of items in the Mokken scale. This extends to all samples analysed here as there were exclusions due to low H_i coefficients in all samples analysed.

Due to the small samples, all item exclusions were made tentatively. In a larger sample, it is likely that some items excluded from these analyses may well have been retained. With this in mind, it is particularly pertinent to take the degree of uncertainty of estimated scalability coefficients when using Mokken's heuristic criteria to determine the strength of scalability when sample size is low [41]. This degree of uncertainty can be assessed and quantified using the SEs of the scalability coefficients. For small samples, i.e. <100, it is important to acknowledge that where SEs are high, the chance of observing a scaling error is high [42]. Where the SE of H is large (for example, 0.08) the probability of the value of H actually being <0.3 is reasonable, which implies that the items within the scale are unscalable [41]. This extends to SEs of item pair and item scalability coefficients. Examining the SEs of item scalability coefficients here suggests there is high likelihood of scaling errors for some items within each of the three IIO hierarchies.

While it is likely that some exclusions due to low H_i values were related to sample size alternatively, these low H_i values of items excluded due to low discrimination (e.g. 'draw overlapping pentagons', 'draw a cube', 'count dot arrays', 'follow written instruction', repetition items and 'word reading') could have been the result of these items assessing some degree of hearing, vision or speech in addition to cognitive ability and therefore not measuring cognitive impairment as succinctly as the rest of the items.

Furthermore, item exclusions could be the result of the similar levels of difficulty of some test items. When several items in a scale assess the same level of the latent trait, IIO cannot be demonstrated [36, 43]. A narrow range of difficulty could result in ISRFs in close proximity to each other, thus making violations of their non-intersection more probable as it is more likely that a particular pattern of response could differ from the expected pattern due to chance alone. While this may be seen as a limitation in the present context and may suggest that some items of the ACE-R could be removed, the similar degree of impairment assessed by some items may be considered an advantage in the context of the test design. As the ACE-R is a screening instrument, it is important to increase the amount of information the instrument reveals about cognitive ability at the level of diagnostic threshold.

Finally, the heterogeneity of the sample could have influenced the number of items retained in the IIO hierarchies. As there were insufficient numbers for separate analyses by dementia diagnosis, we formed three groups and performed all analyses on the three samples. With PCA analysis showing no significant difference between the three groups, the samples were analysed separately for comparison. While this is not ideal, it resulted in three distinct groups for analysis. Future studies should address this limitation by performing aetiology-stratified analyses on various forms of dementia. Mokken scaling of specific aetiologies may result in a greater number of items retained and could help to establish the consistency and rate at which different abilities are lost in various different patient groups.

Furthermore, inconsistencies between individuals can be assessed using person-fit statistics such as PerFit [44]. Person fit methods allow for the identification of unusual response to test items. Detecting unexpected score patterns could have valuable clinical implications and can be useful in improving the interpretation of test scores [45].

High H values (>0.50) of Mokken scales are very seldom reported. The strength of the Mokken scale for the predominantly frontal group (H = 0.52) raises some concern regarding possible violations of LSI. In some cases, elevated H values can reflect LSI violations [46]. As discussed earlier, LSI arises where items are linked, whereby the response to one item is dependent or impossible without prior response to another item. With regard to the ACE-R, some items can be identified as possible sources of LSI violations. For example, 'three-item recall' is linked to the earlier registration of the 3 words in 'three-item registration', and similarly 'name and address recall' and 'recognition' are related to the initial learning of the name and address. There is a strong possibility that these items are not stochastically independent as it is logical that performance in delayed memory recall is predicated on performance on encoding and learning the information to be recalled. However, it is not impossible that a patient could perform better in the recall stage than the learning or repetition stage due to motivational or attentional reasons. Performing Mokken scaling analysis on the ACE-R excluding these 5 items could help determine the degree of LSI violations and the impact of these violations on the scalability of the ACE-R items.

This study of well-phenotyped participants analysed a well-established cognitive test applying novel and robust statistical techniques. The methods applied here, novel in their application to the ACE-R, yielded new and potentially significant findings relevant to both researchers and clinicians. Replication studies of larger samples are required with the present results from this analysis interpreted with caution due to sample size limitations. Mokken scaling analyses applied concurrently with factor analytic methods can provide additional

information, offering prognostic value to clinicians assessing patients. A full neuropsychological assessment is the gold standard in assessing cognitive impairment, but the ACE-R is frequently used not only to determine the degree of cognitive impairment, but also to inspect the extent of subdomain deficits to assist differential diagnosis. Therefore, it is important to note that the item ordering suggests a more complex pattern of decline. While further studies are required to confirm the results and to further delineate the item orderings in sufficiently large distinct diagnostic groups, clinical assessments should expand on merely looking at total scores but should consider the patterns of responses, in particular the order in which the items are failed.

Acknowledgement

The authors are grateful to the patients for their participation. S.M. is supported by a PhD studentship from Alzheimer Scotland. S.M. and J.M.S. are members of the Alzheimer Scotland Dementia Research Centre funded by Alzheimer Scotland. J.M.S. and S.D.S. are members of the University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross-council Lifelong Health and Wellbeing Initiative (MR/K026992/1). J.R.H. is supported by an ARC Federation Fellowship (FF0776229).

Disclosure Statement

The authors report no conflict of interest.

References

1 Wouters H, van Gool WA, Schmand B, Zwinderman AH, Lindeboom R: Three sides of the same coin: measuring global cognitive impairment with the MMSE, ADAS-cog and CAMCOG. Int J Geriat Psychiatry 2010;25:770–779.
2 Reise SP, Ainsworth AT, Haviland MG: Item response theory fundamentals, applications, and promise in psychological research. Curr Dir Psychol 2005;14:95–101.
3 Mathuranath PS, Nestor PJ, Berrios GE, Rakowicz W, Hodges JR: A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia. Neurology 2000;55:1613–1620.
4 Mioshi E, Dawson K, Mitchell J, Arnold R, Hodges JR: The Addenbrooke's Cognitive Examination Revised (ACE-R): a brief cognitive test battery for dementia screening. Int J Geriat Psychiatry 2006;21:1078–1085.
5 Folstein MF, Folstein SE, McHugh PR: 'Mini-mental State'. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–198.
6 Hambleton RK, Swaminathan H: Item Response Theory: Principles and Applications. Boston, Kluwer, 1985.
7 Ligtvoet R, Van der Ark LA, Te Marvelde JM, Sijtsma K: Investigating an invariant item ordering for polytomously scored items. Educ Psychol Meas 2010;70:578–595.
8 Meijer RR, Egberink IJL: Investigating invariant item ordering in personality and clinical scales: some empirical findings and a discussion. Educ Psychol Meas 2012;72:589–607.
9 Sijtsma K, Molenaar IW: Introduction to Nonparametric Item Response Theory. Thousand Oaks, Sage, 2002.
10 Mokken RJ: A Theory and Procedure of Scale Analysis. Berlin, De Gruyter, 1971.
11 Ligtvoet R, van der Ark LA, Bergsma WP, Sijtsma K: Polytomous latent scales for the investigation of the ordering of items. Psychometrik 2011;76a:200–216.
12 Stochl J, Jones PB, Croudace TJ: Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. BMC Med Res Methodol 2012;12:74.
13 Rascovsky K, Hodges JR, Knopman D, Mendez MF, et al: Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 2011;134:2456–2477.
14 McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al: The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:263–269.

15 Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, et al: Classification of primary progressive aphasia and its variants. Neurology 2011;76:1006–1014.
16 Brooks BR, Miller RG, Swash M, Munsat TL: World Federation of Neurology Research Group on Motor Neuron Diseases, El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. Amyotroph Lateral Scler Other Motor Neuron Disord 2000;1:293–299.
17 Neary D, Snowden JS, Gustafson L, Passant U, Stuss D, Black S, et al: Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. Neurology 1998;51:1546–1554.
18 Bentler PM: Comparative fit indexes in structural models. Psychol Bull 1980;107:238–246.
19 Browne MW, Cudeck R: Alternative ways of assessing model fit; in Bollen KA, Long JS (eds): Testing Structural Equation Models. Newbury Park, Sage, 1993, pp 136–162.
20 Kline RB: Principles and Practice of Structural Equation Modelling. New York, Guilford Press, 2005.
21 Arbuckle JL: AMOS (version 18). Chicago, SPSS, 2009.
22 Van der Ark LA: Mokken scale analysis in R. J Stat Softw 2007;20:1–19.
23 Rasch G: Probabilistic Models for Some Intelligence and Attainment Tests. Chicago, University of Chicago, 1960.
24 Guttman L: A basis for scaling qualitative data. Am Soc Rev 1944;9:139–150.
25 Niemoller K, van Schuur W: Stochastic models for unidimensional scaling: Mokken and Rasch; in McKay D, Schoefield N, Whiteley P (eds): Data Analysis and the Social Sciences. London, Francis Pinter, 1983, pp 120–170.
26 Watson R, Wang W, Thompson DR: Violations of local stochastic independence exaggerate scalability in Mokken scaling analysis of the Chinese Mandarin SF-36. Health Qual Life Out 2014;12:1–10.
27 Sijtsma K, Emons WHM, Bouwmeester S, Nyclíček I, Roorda LD: Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization. Quality of Life Scale (WHOQOL-Bref). Qual Life Res 2008;17:275–290.
28 DeJong A, Molenaar IW: An application of Mokken's model for stochastic cumulative scaling in psychiatric research. J Psychiatr Res 1987;21:137–149.
29 Watson R: The Mokken scaling procedure (MSP) applied to the measurement of feeding difficulty in elderly people with dementia. Int J Nurs Stud 1996;33:385–393.
30 Mokken, RJ: Nonparametric models for dichotomous responses; in van der Linden WJ, Hambleton RK: Handbook of Modern Item Response Theory. New York, Springer, 1997, pp 351–367.
31 Nader IW, Tran US, Baranyai P, Voracek M: Investigating dimensionality of Eskin's Attitudes toward Suicide Scale with Mokken scaling and confirmatory factor analysis. Arch Suicide Res 2012;16:226–237.
32 Straat JH: Using Scalability Coefficients and Conditional Association to Assess Monotone Homogeneity. Ridderkerk, Ridderprint BV, 2012.
33 Roorda LD, Scholtes VA, van der Lee JH, Becher J, Dallmeijer AJ: Measuring mobility limitations in children with cerebral palsy: development, scalability, unidimensionality, and internal consistency of the mobility questionnaire, MobQues47. Arch Phys Med Rehabil 2010;91:1194–1209.
34 Meijer RR: A comment on Watson, Deary, and Austin (2007) and Watson, Roberts, Gow, and Deary (2008): How to investigate whether personality items form a hierarchical scale? Pers Individ Dif 2010;48:502–503.
35 Sijtsma K, Meijer RR, van der Ark LA: Mokken Scale Analysis as time goes by: an update for scaling practitioners. Pers Individ Dif 2011;50:31–37.
36 Ligtvoet R: Essays on Invariant Item Ordering. Enschede, Gildeprint Drukkerijen, 2010.
37 Meijer RR, Baneke JJ: Analyzing psychopathology items: a case for nonparametric item response theory modeling. Psychol Methods 2001;9:354–368.
38 Kempen GIJM, Myers AM, Powell L: Hierarchical structure in ADL and IADL: analytical assumptions and applications for clinicians and researchers. J Clin Epidemiol 1995;48:1299–1305.
39 Watson R, Deary IJ, Shipley B: A hierarchy of distress: Mokken scaling of the GHQ-30. Psychol Med 2008;38:575–579.
40 van der Lee JH, Roorda LD, Beckerman H, Lankhorst GJ, Bouter LM: Improving the Action Research Arm test: a unidimensional hierarchical scale. Clin Rehabil 2002;16:646–653.
41 Kuijpers RE, Van der Ark LA, Croon MA: Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. Sociol Methodol 2013;43:42–69.
42 Ringdal K, Ringdal GI, Kaasa S, Bjordal K, Wisløff F, Sundstrøm S, Hjermstad MJ: Assessing the consistency of psychometric properties of the HRQoL scales within the EORTC QLQ-C30 across populations by means of the Mokken Scaling Model. Qual Life Res 1999;8:25–43.
43 Watson R, van der Ark LA, Lin LC, Fieo R, Deary IJ, Meijer RR: Item response theory: how Mokken scaling can be used in clinical practice. J Clin Nurs 2012;21:2736–2746.
44 Tendeiro JN, Tendeiro MJN: Package 'PerFit'. 2014.
45 Meijer RR, Tendeiro JN: The Use of Person-Fit Scores in High-Stakes Educational Testing: How to Use Them and What They Tell Us. Law School Admission Council, Research Report. Newtown, Law School Admission Council, 2014, pp 14–03.
46 Egberink IJ, Meijer RR: An item response theory analysis of Harter's Self-Perception Profile for Children or why strong clinical scales should be distrusted. Assessment 2011;18:201–212.